(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 033 405 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
06.09.2000 Bulletin 2000/36

(21) Application number: 00301439.6

(22) Date of filing: 25.02.2000

(51) Int Cl.7: C12N 15/29, C12N 15/82, C07K 14/415, C12Q 1/68, A01H 5/00

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: 25.02.1999 US 121825 P
27.07.1999 US 145918 P
28.07.1999 US 145951 P
02.08.1999 US 146388 P
02.08.1999 US 146389 P
02.08.1999 US 146386 P
03.08.1999 US 147038 P
04.08.1999 US 147302 P
04.08.1999 US 147204 P
More priorities on the following pages

(83) Declaration under Rule 28(4) EPC (expert solution)

(71) Applicant: Ceres Incorporated
Malibu, CA 90265 (US)

(72) Inventors:
- Alexandrov, Nickolai
Thousand Oaks, CA 91320 (US)
- Brover, Vyacheslav
Calabasas, CA 91302 (US)
- Chen, Xianfeng
Los Angeles, CA 90025 (US)
- Subramanian, Gopalakrishnan
Moorpark, CA 93021 (US)
- Troukhan, Maxim E.
South Pasadena, CA 91030 (US)
- Zheng, Liansheng
Creve Coeur, MO 63141 (US)
- Dumas, J.
, (US)

(74) Representative:
Bannerman, David Gardner et al
Withers & Rogers,
Goldings House,
2 Hays Lane
London SE1 2HW (GB)

Remarks:
THE COMPLETE DOCUMENT INCLUDING REFERENCE TABLES AND THE SEQUENCE LISTING IS AVAILABLE ON CD-ROM FROM THE EUROPEAN PATENT OFFICE, VIENNA SUB-OFFICE.

(54) Sequence-determined DNA fragments and corresponding polypeptides encoded thereby

(57) The present invention provides DNA molecules that constitute fragments of the genome of a plant, and polypeptides encoded thereby. The DNA molecules are useful for specifying a gene product in cells, either as a promoter or as a protein coding sequence or as an UTR or as a 3' termination sequence, and are also useful in controlling the behavior of a gene in the chromosome, in controlling the expression of a gene or as tools for genetic mapping, recognizing or isolating identical or related DNA fragments, or identification of a particular individual organism, or for clustering of a group of organisms with a common trait.

[0]Arabidopsis DNA is used in the present experiment, but the procedure is a general one.

EP 1 033 405 A2

(30) Priority: Continued from first page

05.08.1999 US 147260 P
05.08.1999 US 147192 P
06.08.1999 US 147303 P
06.08.1999 US 147416 P
09.08.1999 US 147493 P
09.08.1999 US 147935 P
10.08.1999 US 148171 P
11.08.1999 US 148319 P
12.08.1999 US 148341 P
13.08.1999 US 148565 P
13.08.1999 US 148684 P
16.08.1999 US 149368 P
17.08.1999 US 149175 P
18.08.1999 US 149426 P
20.08.1999 US 149722 P
20.08.1999 US 149929 P
20.08.1999 US 149723 P
23.08.1999 US 149902 P
23.08.1999 US 149930 P
25.08.1999 US 150566 P
26.08.1999 US 150884 P
27.08.1999 US 151065 P
27.08.1999 US 151066 P
27.08.1999 US 151080 P
30.08.1999 US 151303 P
31.08.1999 US 151438 P
01.09.1999 US 151930 P
07.09.1999 US 152363 P
10.09.1999 US 153070 P
13.09.1999 US 153758 P
15.09.1999 US 154018 P
16.09.1999 US 154039 P
20.09.1999 US 154779 P
22.09.1999 US 155139 P
23.09.1999 US 155486 P
24.09.1999 US 155659 P
28.09.1999 US 156458 P
29.09.1999 US 156596 P
04.10.1999 US 157117 P
05.10.1999 US 157753 P
06.10.1999 US 157865 P
07.10.1999 US 158029 P
08.10.1999 US 158232 P
12.10.1999 US 158369 P
13.10.1999 US 159294 P
13.10.1999 US 159295 P
13.10.1999 US 159293 P
14.10.1999 US 159638 P
14.10.1999 US 159637 P
14.10.1999 US 159329 P
14.10.1999 US 159331 P
14.10.1999 US 159330 P
18.10.1999 US 159584 P
21.10.1999 US 160815 P
21.10.1999 US 160767 P
21.10.1999 US 160768 P
21.10.1999 US 160741 P
21.10.1999 US 160770 P
21.10.1999 US 160814 P
22.10.1999 US 160981 P
22.10.1999 US 160980 P
22.10.1999 US 160989 P
25.10.1999 US 161405 P
25.10.1999 US 161404 P
25.10.1999 US 161406 P
26.10.1999 US 161361 P
26.10.1999 US 161360 P
26.10.1999 US 161359 P
28.10.1999 US 161920 P
28.10.1999 US 161992 P
28.10.1999 US 161993 P
29.10.1999 US 162143 P
29.10.1999 US 162142 P
29.10.1999 US 162228 P
01.11.1999 US 162895 P
01.11.1999 US 162891 P
01.11.1999 US 162894 P
02.11.1999 US 163093 P
02.11.1999 US 163092 P
02.11.1999 US 163091 P
03.11.1999 US 163249 P
03.11.1999 US 163248 P
03.11.1999 US 163281 P
04.11.1999 US 163380 P
04.11.1999 US 163381 P
04.11.1999 US 163379 P
08.11.1999 US 164151 P
08.11.1999 US 164150 P
08.11.1999 US 164146 P
09.11.1999 US 164260 P
09.11.1999 US 164259 P
10.11.1999 US 164548 P
10.11.1999 US 164317 P
10.11.1999 US 164321 P
10.11.1999 US 164318 P
10.11.1999 US 164544 P
10.11.1999 US 164545 P
10.11.1999 US 164319 P
12.11.1999 US 164870 P
12.11.1999 US 164959 P
12.11.1999 US 164962 P
12.11.1999 US 164960 P
12.11.1999 US 164871 P
12.11.1999 US 164961 P
15.11.1999 US 164927 P
15.11.1999 US 164929 P
15.11.1999 US 164926 P
16.11.1999 US 165669 P
16.11.1999 US 165671 P
16.11.1999 US 165661 P
17.11.1999 US 165919 P
17.11.1999 US 165918 P
17.11.1999 US 165911 P
18.11.1999 US 166158 P
18.11.1999 US 166157 P

(Cont. next page)

18.11.1999 US 166173 P
19.11.1999 US 166412 P
19.11.1999 US 166419 P
19.11.1999 US 166411 P
22.11.1999 US 166733 P
22.11.1999 US 166750 P
23.11.1999 US 167362 P
24.11.1999 US 167382 P
24.11.1999 US 167233 P
24.11.1999 US 167234 P
24.11.1999 US 167235 P
30.11.1999 US 167904 P
30.11.1999 US 167908 P
30.11.1999 US 167902 P
01.12.1999 US 168232 P
01.12.1999 US 168233 P
01.12.1999 US 168231 P
02.12.1999 US 168546 P
02.12.1999 US 168549 P
02.12.1999 US 168548 P
03.12.1999 US 168673 P
03.12.1999 US 168675 P
03.12.1999 US 168674 P
07.12.1999 US 169278 P
07.12.1999 US 169302 P
07.12.1999 US 169298 P
08.12.1999 US 169692 P
08.12.1999 US 169691 P
16.12.1999 US 171107 P
16.12.1999 US 171098 P
16.12.1999 US 171114 P
19.01.2000 US 176866 P
19.01.2000 US 176867 P
19.01.2000 US 176910 P
26.01.2000 US 178166 P
27.01.2000 US 178547 P
27.01.2000 US 177666 P
27.01.2000 US 178546 P
27.01.2000 US 178544 P
27.01.2000 US 178545 P
28.01.2000 US 178755 P
28.01.2000 US 178754 P
01.02.2000 US 179395 P
01.02.2000 US 179388 P
03.02.2000 US 180039 P
03.02.2000 US 180139 P
04.02.2000 US 180207 P
04.02.2000 US 180206 P
07.02.2000 US 180695 P
07.02.2000 US 180696 P
09.02.2000 US 181228 P
09.02.2000 US 181214 P
10.02.2000 US 181476 P
10.02.2000 US 181551 P
15.02.2000 US 182477 P
15.02.2000 US 182516 P
15.02.2000 US 182512 P
15.02.2000 US 182478 P
17.02.2000 US 183165 P
17.02.2000 US 183166 P
27.07.1999 US 145913 P
05.03.1999 US 123180 P
09.03.1999 US 123548 P
23.03.1999 US 125788 P
25.03.1999 US 126264 P
29.03.1999 US 126785 P
01.04.1999 US 127462 P
06.04.1999 US 128234 P
08.04.1999 US 128714 P
16.04.1999 US 129845 P
19.04.1999 US 130077 P
21.04.1999 US 130449 P
23.04.1999 US 130891 P
23.04.1999 US 130510 P
28.04.1999 US 131449 P
30.04.1999 US 132407 P
30.04.1999 US 132048 P
04.05.1999 US 132484 P
05.05.1999 US 132485 P
06.05.1999 US 132487 P
06.05.1999 US 132486 P
07.05.1999 US 132863 P
11.05.1999 US 134256 P
14.05.1999 US 134221 P
14.05.1999 US 134218 P
14.05.1999 US 134370 P
14.05.1999 US 134219 P
18.05.1999 US 134768 P
19.05.1999 US 134941 P
20.05.1999 US 135124 P
21.05.1999 US 135353 P
24.05.1999 US 135629 P
25.05.1999 US 136021 P
27.05.1999 US 136392 P
28.05.1999 US 136782 P
01.06.1999 US 137222 P
03.06.1999 US 137528 P
04.06.1999 US 137502 P
07.06.1999 US 137724 P
08.06.1999 US 138094 P
10.06.1999 US 138540 P
10.06.1999 US 138847 P
14.06.1999 US 139119 P
16.06.1999 US 139452 P
16.06.1999 US 139453 P
17.06.1999 US 139492 P
18.06.1999 US 139461 P
18.06.1999 US 139750 P
18.06.1999 US 139463 P
18.06.1999 US 139457 P
18.06.1999 US 139459 P
18.06.1999 US 139462 P
18.06.1999 US 139455 P
18.06.1999 US 139458 P
18.06.1999 US 139454 P
18.06.1999 US 139456 P

(Cont. next page)

18.06.1999	US 139460 P
18.06.1999	US 139763 P
21.06.1999	US 139817 P
22.06.1999	US 139899 P
23.06.1999	US 140354 P
23.06.1999	US 140353 P
24.06.1999	US 140695 P
28.06.1999	US 140823 P
29.06.1999	US 140991 P
30.06.1999	US 141287 P
01.07.1999	US 142154 P
01.07.1999	US 141842 P
02.07.1999	US 142055 P
06.07.1999	US 142390 P
08.07.1999	US 142803 P
09.07.1999	US 142920 P
12.07.1999	US 142977 P
13.07.1999	US 143542 P
14.07.1999	US 143624 P
15.07.1999	US 144005 P
16.07.1999	US 144085 P
16.07.1999	US 144086 P
19.07.1999	US 144333 P
19.07.1999	US 144335 P
19.07.1999	US 144325 P
19.07.1999	US 144334 P
19.07.1999	US 144332 P
19.07.1999	US 144331 P
20.07.1999	US 144884 P
20.07.1999	US 144352 P
20.07.1999	US 144632 P
21.07.1999	US 144814 P
21.07.1999	US 145086 P
21.07.1999	US 145088 P
22.07.1999	US 145192 P
22.07.1999	US 145085 P
22.07.1999	US 145089 P
22.07.1999	US 145087 P
23.07.1999	US 145145 P
23.07.1999	US 145224 P
23.07.1999	US 145218 P
26.07.1999	US 145276 P
27.07.1999	US 145919 P

Description

FIELD OF THE INVENTION

[0001] The present invention relates to isolated polynucleotides that represent a complete gene, or a fragment thereof, that is expressed. In addition, the present invention relates to the polypeptide or protein corresponding to the coding sequence of these polynucleotides. The present invention also relates to isolated polynucleotides that represent regulatory regions of genes. The present invention also relates to isolated polynucleotides that represent untranslated regions of genes. The present invention further relates to the use of these isolated polynucleotides and polypeptides and proteins.

DESCRIPTION OF THE RELATED ART

[0002] Efforts to map and sequence the genome of a number of organisms are in progress; a few complete genome sequences, for example those of *E. coli* and *Saccharomyces cerevisiae* are known (Blattner et al., *Science* 277: 1453 (1997); Goffeau et al., *Science* 274:546 (1996)). The complete genome of a multicellular organism, *C. elegans*, has also been sequenced (See, the *C. elegans* Sequencing Consortium, *Science* 282:2012 (1998)). To date, no complete genome of a plant has been sequenced, nor has a complete cDNA complement of any plant been sequenced.

SUMMARY OF THE INVENTION

[0003] The present invention comprises polynucleotides, such as complete cDNA sequences and/or sequences of genomic DNA encompassing complete genes, fragments of genes, and/or regulatory elements of genes and/or regions with other functions and/or intergenic regions, hereinafter collectively referred to as Sequence-Determined DNA Fragments (SDFs), from different plant species, particularly corn, wheat, soybean, rice and *Arabidopsis thaliana*, and other plants and or mutants, variants, fragments or fusions of said SDFs and polypeptides or proteins derived therefrom. In some instances, the SDFs span the entirety of a protein-coding segment. In some instances, the entirety of an mRNA is represented. Other objects of the invention that are also represented by SDFs of the invention are control sequences, such as, but not limited to, promoters. Complements of any sequence of the invention are also considered part of the invention.

[0004] Other objects of the invention are polynucleotides comprising exon sequences, polynucleotides comprising intron sequences, polynucleotides comprising introns together with exons, intron/exon junction sequences, 5' untranslated sequences, and 3' untranslated sequences of the SDFs of the present invention. Polynucleotides representing the joinder of any exons described herein, in any arrangement, for example, to produce a sequence encoding any desirable amino acid sequence are within the scope of the invention.

[0005] The present invention also resides in probes useful for isolating and identifying nucleic acids that hybridize to an SDF of the invention. The probes can be of any length, but more typically are 12-2000 nucleotides in length; more typically, 15 to 200 nucleotides long; even more typically, 18 to 100 nucleotides long.

[0006] Yet another object of the invention is a method of isolating and/or identifying nucleic acids using the following steps:

(a) contacting a probe of the instant invention with a polynucleotide sample under conditions that permit hybridization and formation of a polynucleotide duplex; and
(b) detecting and/or isolating the duplex of step (a).

[0007] The conditions for hybridization can be from low to moderate to high stringency conditions. The sample can include a polynucleotide having a sequence unique in a plant genome. Probes and methods of the invention are useful, for example, without limitation, for mapping of genetic traits and/or for positional cloning of a desired fragment of genomic DNA.

[0008] Probes and methods of the invention can also be used for detecting alternatively spliced messages within a species. Probes and methods of the invention can further be used to detect or isolate related genes in other plant species using genomic DNA (gDNA) and/or cDNA libraries. In some instances, especially when longer probes and low to moderate stringency hybridization conditions are used, the probe will hybridize to a plurality of cDNA and/or gDNA sequences of a plant. This approach is useful for isolating representatives of gene families which are identifiable by possession of a common functional domain in the gene product or which have common cis-acting regulatory sequences. This approach is also useful for identifying orthologous genes from other organisms.

[0009] The present invention also resides in constructs for modulating the expression of the genes comprised of all or a fragment of an SDF. The constructs comprise all or a fragment of the expressed SDF, or of a complementary

sequence. Examples of constructs include ribozymes comprising RNA encoded by an SDF or by a sequence complementary thereto, antisense constructs, constructs comprising coding regions or parts thereof, constructs comprising promoters, introns, untranslated regions, scaffold attachment regions, methylating regions, enhancing or reducing regions, DNA and chromatin conformation modifying sequences, etc. Such constructs can be constructed using viral, plasmid, bacterial artificial chromosomes (BACs), plasmid artificial chromosomes (PACs), autonomous plant plasmids, plant artificial chromosomes or other types of vectors and exist in the plant as autonomous replicating sequences or as DNA integrated into the genome. When inserted into a host cell the construct is, preferably, functionally integrated with, or operatively linked to, a heterologous polynucleotide. For instance, a coding region from an SDF might be operably linked to a promoter that is functional in a plant.

[0010] The present invention also resides in host cells, including bacterial or yeast cells or plant cells, and plants that harbor constructs such as described above. Another aspect of the invention relates to methods for modulating expression of specific genes in plants by expression of the coding sequence of the constructs, by regulation of expression of one or more endogenous genes in a plant or by suppression of expression of the polynucleotides of the invention in a plant. Methods of modulation of gene expression include without limitation (1) inserting into a host cell additional copies of a polynucleotide comprising a coding sequence; (2) modulating an endogenous promoter in a host cell; (3) inserting antisense or ribozyme constructs into a host cell and (4) inserting into a host cell a polynucleotide comprising a sequence encoding a variant, fragment, or fusion of the native polypeptides of the instant invention.

BRIEF DESCRIPTION OF THE TABLES

[0011] The sequences of exemplary SDFs and polypeptides corresponding to the coding sequences of the instant invention are described in Reference Tables 1 and 2, REF Tables 1 and 2"; and in Sequence Tables 1 and 2, SEQ Tables 1 and 2." The REF Tables refer to a number of Maximum Length Sequences" or MLS." Each MLS corresponds to the longest cDNA obtained, either by cloning or by the prediction from genomic sequence. The sequence of the MLS is the cDNA sequence as described in the Av subsection of the REF Tables.

[0012] The REF Table includes the following information relating to each MLS:

I. cDNA Sequence

- A. 5' UTR
- B. Coding Sequence
- C. 3' UTR

II. Genomic Sequence

- A. Exons
- B. Introns
- C. Promoters

III. Link of cDNA Sequences to Clone IDs
IV. Multiple Transcription Start Sites
V. Polypeptide Sequences

- A. Signal Peptide
- B. Domains
- C. Related Polypeptides

VI. Related Polynucleotide Sequences

I. cDNA SEOUENCE

[0013] The REF Tables indicate which sequence in the SEQ Tables represents the sequence of each MLS. The MLS sequence can comprise 5' and 3' UTR as well as coding sequences. In addition, specific cDNA clone numbers also are included in the REF Tables when the MLS sequence relates to a specific cDNA clone.

A. 5' UTR

[0014] The location of the 5' UTR can be determined by comparing the most 5' MLS sequence with the corresponding

genomic sequence as indicated in the REF Tables. The sequence that matches, beginning at any of the transcriptional start sites and ending at the last nucleotide before any of the translational start sites corresponds to the 5' UTR.

B. Coding Region

[0015] The coding region is the sequence in any open reading frame found in the MLS. Coding regions of interest are indicated in the Poly P SEQ subsection of the REF Tables.

C. 3' UTR

[0016] The location of the 3' UTR can be determined by comparing the most 3' MLS sequence with the corresponding genomic sequence as indicated in the REF Tables. The sequence that matches, beginning at the translational stop site and ending at the last nucleotide of the MLS corresponds to the 3' UTR.

II. GENOMIC SEQUENCE

[0017] Further, the REF Tables indicate the specific gi* number of the genomic sequence if the sequence resides in a public databank. For each genomic sequence, the REF Tables indicate which regions are included in the MLS. These regions can include the 5' and 3' UTRs as well as the coding sequence of the MLS. See, for example, the scheme below:

```
                  Region 1                 Region 2             Region 3
           ___________________                    ___________      _________________
---------| 5' UTR   | Exon |---------| Exon |--------| Exon | 3' UTR |-------
    ^                              ^                         ^                      ^
    |                ^             |                  |             ^               |
Promoter             |           Intron             Intron          |
              Translational                                     Stop Codon
               Start Site
```

[0018] The REF Tables report the first and last base of each region that are included in an MLS sequence. An example is shown below:

gi No. 47000:
37102 ... 37497
37593 ... 37925

The numbers indicate that the MLS contains the following sequences from two regions of gi No. 47000; a first region including bases 37102-37497, and a second region including bases 37593-37925.

A. EXON SEQUENCES

[0019] The location of the exons can be determined by comparing the sequence of the regions from the genomic sequences with the corresponding MLS sequence as indicated by the REF Tables.

I. INITIAL EXON

[0020] To determine the location of the initial exon, information from the

(1) polypeptide sequence section;
(2) cDNA polynucleotide section: and
(3) the genomic sequence section

of the REF Tables are used. First, the polypeptide section will indicate where the translational start site is located in the MLS sequence. The MLS sequence can be matched to the genomic sequence that corresponds to the MLS. Based on the match between the MLS and corresponding genomic sequences , the location of the translational start site can be determined in one of the regions of the genomic sequence. The location of this translational start site is the start of the first exon.

[0021] Generally, the last base of the exon of the corresponding genomic region, in which the translational start site

was located, will represent the end of the initial exon. In some cases, the initial exon will end with a stop codon, when the initial exon is the only exon.

[0022] In the case when sequences representing the MLS are in the positive strand of the corresponding genomic sequence, the last base will be a larger number than the first base. When the sequences representing the MLS are in the negative strand of the corresponding genomic sequence, then the last base will be a smaller number than the first base.

ii. INTERNAL EXONS

[0023] Except for the regions that comprise the 5' and 3' UTRs, initial exon, and terminal exon, the remaining genomic regions that match the MLS sequence are the internal exons. Specifically, the bases defining the boundaries of the remaining regions also define the intron/exon junctions of the internal exons.

iii. TERMINAL EXON

[0024] As with the initial exon, the location of the terminal exon is determined with information from the

(1) polypeptide sequence section;
(2) cDNA polynucleotide section; and
(3) the genomic sequence section

of the REF Tables. The polypeptide section will indicate where the stop codon is located in the MLS sequence. The MLS sequence can be matched to the corresponding genomic sequence. Based on the match between MLS and corresponding genomic sequences, the location of the stop codon can be determined in one of the regions of the genomic sequence. The location of this stop codon is the end of the terminal exon. Generally, the first base of the exon of the corresponding genomic region that matches the cDNA sequence, in which the stop codon was located, will represent the beginning of the terminal exon. In some cases, the translational start site will represent the start of the terminal exon, which will be the only exon.

[0025] In the case when the MLS sequences are in the positive strand of the corresponding genomic sequence, the last base will be a larger number than the first base. When the MLS sequences are in the negative strand of the corresponding genomic sequence, then the last base will be a smaller number than the first base.

B. INTRON SEQUENCES

[0026] In addition, the introns corresponding to the MLS are defined by identifying the genomic sequence located between the regions where the genomic sequence comprises exons. Thus, introns are defined as starting one base downstream of a genomic region comprising an exon, and end one base upstream from a genomic region comprising an exon.

C. PROMOTER SEQUENCES

[0027] As indicated below, promoter sequences corresponding to the MLS are defined as sequences upstream of the first exon; more usually, as sequences upstream of the first of multiple transcription start sites; even more usually as sequences about 2,000 nucleotides upstream of the first of multiple transcription start sites.

III. LINK of cDNA SEQUENCES to CLONE IDs

[0028] As noted above, the REF tables identify the cDNA clone(s) that relate to each MLS. The MLS sequence can be longer than the sequences included in the cDNA clones. In such a case, the REF table indicates the region of the MLS that is included in the clone. If either the 5' or 3' termini of the cDNA clone sequence is the same as the MLS sequence, no mention will be made.

IV. Multiple Transcription Start Sites

[0029] Initiation of transcription can occur at a number of sites of the gene. The REF tables indicate the possible multiple transcription sites for each gene. In the REF tables, the location of the transcription start sites can be either a positive or negative number. The positions indicated by positive numbers refer to the transcription start sites as located in the MLS sequence. The negative numbers indicate the transcription start site within the genomic sequence

that corresponds to the MLS.

[0030] To determine the location of the transcription start sites with the negative numbers, the MLS sequence is aligned with the corresponding genomic sequence. In the instances when a public genomic sequence is referenced, the relevant corresponding genomic sequence can be found by direct reference to the nucleotide sequence indicated by the gi" number shown in the public genomic DNA section of the REF tables. When the position is a negative number, the transcription start site is located in the corresponding genomic sequence upstream of the base that matches the beginning of the MLS sequence in the alignment. The negative number is relative to the first base of the MLS sequence which matches the genomic sequence corresponding to the relevant gi" number.

[0031] In the instances when no public genomic DNA is referenced, the relevant nucleotide sequence for alignment is the nucleotide sequence associated with the amino acid sequence designated by gi" number of the later PolyP SEQ subsection.

V. Polypeptide Sequences

[0032] The PolyP SEQ subsection lists SEQ ID NOs and Ceres SEQ ID NO for polypeptide sequences corresponding to the coding sequence of the MLS sequence and the location of the translational start site with the coding sequence of the MLS sequence.

[0033] The MLS sequence can have multiple translational start sites and can be capable of producing more than one polypeptide sequence.

A. Signal Peptide

[0034] The REF Tables also indicate in subsection (B) the cleavage site of the putative signal peptide of the polypeptide corresponding to the coding sequence of the MLS sequence. Typically, signal peptide coding sequences comprise a sequence encoding the first residue of the polypeptide to the cleavage site residue.

B. Domains

[0035] Subsection (C) provides information regarding identified domains (where present) within the polypeptide and (where present) a name for the polypeptide domain.

C. Related Polypeptides

[0036] Subsection (Dp) provides (where present) information concerning amino acid sequences that are found to be related and have some percentage of sequence identity to the polypeptide sequences of REF and SEQ TABLES 1 AND 2. These related sequences are identified by a gi" number.

VI. Related Polynucleotide Sequences

[0037] Subsection (Dn) provides polynucleotide sequences (where present) that are related to and have some percentage of sequence identity to the MLS or corresponding genomic sequence.

Abbreviation	Description
Max Len. Seq..	Maximum Length Sequence
rel to	Related to
Clone Ids	Clone ID numbers
Pub gDNA	Public Genomic DNA
gi No.	gi number
Gen. seq. in cDNA	Genomic Sequence in cDNA (Each region for a single gene prediction is listed on a separate line.
	In the case of multiple gene predictions, the group of regions relating to a single prediction are separated by a blank line)
(Ac) cDNA SEQ	cDNA sequence

(continued)

Abbreviation	Description
- Pat. Appln. SEQ ID NO	Patent Application SEQ ID NO:
- Ceres SEQ ID NO: 1673877	Ceres SEQ ID NO:
- SEQ # w. TSS	Location within the cDNA sequence, SEQ ID NO:, of Transcription Start Sites which are listed below
- Clone ID #: # -> #	Clone ID comprises bases # to # of the cDNA Sequence
PolyP SEQ	Polypeptide Sequence
- Pat. Appln. SEQ ID NO:	Patent Application SEQ ID NO:
- Ceres SEQ ID NO	Ceres SEQ ID NO:
- Loc. SEQ ID NO: @ nt.	Location of translational start site in cDNA of SEQ ID NO: at nucleotide number
(C) Pred. PP Nom. & Annot.	Nomination and Annotation of Domains within Predicted Polypeptide(s)
- (Title)	Name of Domain
- Loc. SEQ ID NO #: # -> # aa.	Location of the domain within the polypeptide of SEQ ID NO: from # to # amino acid residues.
(Dp) Rel. AA SEQ	Related Amino Acid Sequences
- Align. NO	Alignment number
- gi No	Gi number
- Desp.	Description
- % Idnt.	Percent identity
- Align. Len.	Alignment Length
- Loc. SEQ ID NO: # -> # aa	Location within SEQ ID NO: from # to # amino acid residue.

DETAILED DESCRIPTION OF THE INVENTION

[0038] The invention relates to (I) polynucleotides and methods of use thereof, such as

IA. Probes, Primers and Substrates;
IB. Methods of Detection and Isolation;

B.1. Hybridization;
B.2. Methods of Mapping;
B.3. Southern Blotting;
B.4. Isolating cDNA from Related Organisms;
B.5. Isolating and/or Identifying Orthologous Genes

IC. Methods of Inhibiting Gene Expression

C.1. Antisense
C.2. Ribozyme Constructs;
C.3. Chimeraplasts;
C.4 Co-Suppression;
C.5. Transcriptional Silencing
C.6. Other Methods to Inhibit Gene Expression

ID. Methods of Functional Analysis;
IE. Promoter Sequences and Their Use;

IF. UTRs and/or Intron Sequences and Their Use; and
IG. Coding Sequences and Their Use.

[0039] The invention also relates to (II) polypeptides and proteins and methods of use thereof, such as

IIA. Native Polypeptides and Proteins

A.1 Antibodies
A.2 In Vitro Applications

IIB. Polypeptide Variants, Fragments and Fusions

B.1 Variants
B.2 Fragments
B.3 Fusions

[0040] The invention also includes (III) methods of modulating polypeptide production, such as

IIIA. Suppression

A.1 Antisense
A.2 Ribozymes
A.3 Co-suppression
A.4 Insertion of Sequences into the Gene to be Modulated
A.5 Promoter Modulation
A.6 Expression of Genes containing Dominant-Negative Mutations

IIIB. Enhanced Expression

B.1 Insertion of an Exogenous Gene
B.2 Promoter Modulation

[0041] The invention further concerns (IV) gene constructs and vector construction, such as

IVA. Coding Sequences
IVB. Promoters
IVC. Signal Peptides

[0042] The invention still further relates to
V Transformation Techniques

Definitions

[0043] Allelic variant An allelic variant" is an alternative form of the same SDF, which resides at the same chromosomal locus in the organism. Allelic variations can occur in any portion of the gene sequence, including regulatory regions. Allelic variants can arise by normal genetic variation in a population. Allelic variants can also be produced by genetic engineering methods. An allelic variant can be one that is found in a naturally occurring plant, including a cultivar or ecotype. An allelic variant may or may not give rise to a phenotypic change, and may or may not be expressed. An allele can result in a detectable change in the phenotype of the trait represented by the locus. A phenotypically silent allele can give rise to a product.

[0044] Alternatively spliced messages Within the context of the current invention, alternatively spliced messages" refers to mature mRNAs originating from a single gene with variations in the number and/or identity of exons, introns and/or intron-exon junctions.

[0045] Chimeric The term chimeric" is used to describe genes, as defined supra, or contructs wherein at least two of the elements of the gene or construct, such as the promoter and the coding sequence and/or other regulatory sequences and/or filler sequences and/or complements thereof, are heterologous to each other.

[0046] Constitutive Promoter: Promoters referred to herein as "constitutive promoters" actively promote transcription under most, but not necessarily all, environmental conditions and states of development or cell differentiation. Examples

of constitutive promoters include the cauliflower mosaic virus (CaMV) 35S transcript initiation region and the 1' or 2' promoter derived from TDNA of *Agrobacterium tumefaciens,* and other transcription initiation regions from various plant genes, such as the maize ubiquitin-1 promoter, known to those of skill.

[0047] Coordinately Expressed: The term coordinately expressed," as used in the current invention, refers to genes that are expressed at the same or a similar time and/or stage and/or under the same or similar environmental conditions.

[0048] Domain: Domains are fingerprints or signatures that can be used to characterize protein families and/or parts of proteins. Such fingerprints or signatures can comprise conserved (1) primary sequence, (2) secondary structure, and/or (3) three-dimensional conformation. Generally, each domain has been associated with either a family of proteins or motifs. Typically, these families and/or motifs have been correlated with specific *in-vitro* and/or *in-vivo* activities. A domain can be any length, including the entirety of the sequence of a protein. Detailed descriptions of the domains, associated families and motifs, and correlated activities of the polypeptides of the instant invention are described below. Usually, the polypeptides with designated domain(s) can exhibit at least one activity that is exhibited by any polypeptide that comprises the same domain(s).

[0049] Endogenous The term endogenous," within the context of the current invention refers to any polynucleotide, polypeptide or protein sequence which is a natural part of a cell or organisms regenerated from said cell.

[0050] Exogenous Exogenous," as referred to within, is any polynucleotide, polypeptide or protein sequence, whether chimeric or not, that is initially or subsequently introduced into the genome of an individual host cell or the organism regenerated from said host cell by any means other than by a sexual cross. Examples of means by which this can be accomplished are described below, and include *Agrobacterium*-mediated transformation (of dicots - *e.g.* Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983); of monocots, representative papers are those by Escudero et al., *Plant J.* 10:355 (1996), Ishida et al., *Nature Biotechnology* 14:745 (1996), May et al., *Bio/Technology* 13:486 (1995)), biolistic methods (Armaleo et al., *Current Genetics* 17:97 1990)), electroporation, *in planta* techniques, and the like. Such a plant containing the exogenous nucleic acid is referred to here as a T_0 for the primary transgenic plant and T_1 for the first generation. The term exogenous" as used herein is also intended to encompass inserting a naturally found element into a non-naturally found location.

[0051] Filler sequence: As used herein, filler sequence" refers to any nucleotide sequence that is inserted into DNA construct to evoke a particular spacing between particular components such as a promoter and a coding region and may provide an additional attribute such as a restriction enzyme site.

[0052] Gene: The term gene," as used in the context of the current invention, encompasses all regulatory and coding sequence contiguously associated with a single hereditary unit with a genetic function (see SCHEMATIC 1). Genes can include non-coding sequences that modulate the genetic function that include, but are not limited to, those that specify polyadenylation, transcriptional regulation, DNA conformation, chromatin conformation, extent and position of base methylation and binding sites of proteins that control all of these. Genes comprised of exons" (coding sequences), which may be interrupted by introns" (non-coding sequences), encode proteins. A gene's genetic function may require only RNA expression or protein production, or may only require binding of proteins and/or nucleic acids without associated expression. In certain cases, genes adjacent to one another may share sequence in such a way that one gene will overlap the other. A gene can be found within the genome of an organism, artificial chromosome, plasmid, vector, etc., or as a separate isolated entity.

[0053] Gene Family: Gene family" is used in the current invention to describe a group of functionally related genes, each of which encodes a separate protein.

[0054] Heterologous sequences: Heterologous sequences" are those that are not operatively linked or are not contiguous to each other in nature. For example, a promoter from corn is considered heterologous to an *Arabidopsis* coding region sequence. Also, a promoter from a gene encoding a growth factor from corn is considered heterologous to a sequence encoding the corn receptor for the growth factor. Regulatory element sequences, such as UTRs or 3' end termination sequences that do not originate in nature from the same gene as the coding sequence originates from, are considered heterologous to said coding sequence. Elements operatively linked in nature and contiguous to each other are not heterologous to each other. On the other hand, these same elements remain operativley linked but become heterologous if other filler sequence is placed between them. Thus, the promoter and coding sequences of a corn gene expressing an amino acid transporter are not heterologous to each other, but the promoter and coding sequence of a corn gene operatively linked in a novel manner are heterologous.

[0055] Homologous gene In the current invention, homologous gene" refers to a gene that shares sequence similarity with the gene of interest. This similarity may be in only a fragment of the sequence and often represents a functional domain such as, examples including without limitation a DNA binding domain, a domain with tyrosine kinase activity, or the like. The functional activities of homologous genes are not necessarily the same.

[0056] Inducible Promoter An inducible promoter" in the context of the current invention refers to a promoter which is regulated under certain conditions, such as light, chemical concentration, protein concentration, conditions in an organism, cell, or organelle, etc. A typical example of an inducible promoter, which can be utilized with the polynu-

cleotides of the present invention, is PARSK1, the promoter from the *Arabidopsis* gene encoding a serine-threonine kinase enzyme, and which promoter is induced by dehydration, abscissic acid and sodium chloride (Wang and Goodman, *Plant J.* 8:37 (1995)) Examples of environmental conditions that may affect transcription by inducible promoters include anaerobic conditions, elevated temperature, or the presence of light.

[0057] Intergenic region Intergenic region," as used in the current invention, refers to nucleotide sequence occurring in the genome that separates adjacent genes.

[0058] Mutant gene In the current invention, mutant" refers to a heritable change in DNA sequence at a specific location. Mutants of the current invention may or may not have an associated identifiable function when the mutant gene is transcribed.

[0059] Orthologous Gene In the current invention orthologous gene" refers to a second gene that encodes a gene product that performs a similar function as the product of a first gene. The orthologous gene may also have a degree of sequence similarity to the first gene. The orthologous gene may encode a polypeptide that exhibits a degree of sequence similarity to a polypeptide corresponding to a first gene. The sequence similarity can be found within a functional domain or along the entire length of the coding sequence of the genes and/or their corresponding polypeptides.

[0060] Percentage of sequence identity "Percentage of sequence identity," as used herein, is determined by comparing two optimally aligned sequences over a comparison window, where the fragment of the polynucleotide or amino acid sequence in the comparison window may comprise additions or deletions (e.g., gaps or overhangs) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and *Waterman Add. APL. Math.* 2:482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48:443 (1970), by the search for similarity method of Pearson and Lipman *Proc. Natl. Acad Sci. (USA)* 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT, BLAST, PASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, WI), or by inspection. Given that two sequences have been identified for comparison, GAP and BESTFIT are preferably employed to determine their optimal alignment. Typically, the default values of 5.00 for gap weight and 0.30 for gap weight length are used. The term "substantial sequence identity" between polynucleotide or polypeptide sequences refers to polynucleotide or polypeptide comprising a sequence that has at least 80% sequence identity, preferably at least 85%, more preferably at least 90% and most preferably at least 95%, even more preferably, at least 96%, 97%, 98% or 99% sequence identity compared to a reference sequence using the programs.

[0061] Plant Promoter A plant promoter" is a promoter capable of initiating transcription in plant cells and can drive or facilitate transcription of a fragment of the SDF of the instant invention or a coding sequence of the SDF of the instant invention. Such promoters need not be of plant origin. For example, promoters derived from plant viruses, such as the CaMV35S promoter or from *Agrobacterium tumefaciens* such as the T-DNA promoters, can be plant promoters. A typical example of a plant promoter of plant origin is the maize ubiquitin-1 (ubi-1)promoter known to those of skill.

[0062] Promoter: The term "promoter," as used herein, refers to a region of sequence determinants located upstream from the start of transcription of a gene and which are involved in recognition and binding of RNA polymerase and other proteins to initiate and modulate transcription. A basal promoter is the minimal sequence necessary for assembly of a transcription complex required for transcription initiation. Basal promoters frequently include a TATA box" element usually located between 15 and 35 nucleotides upstream from the site of initiation of transcription. Basal promoters also sometimes include a CCAAT box" element (typically a sequence CCAAT) and/or a GGGCG sequence, usually located between 40 and 200 nucleotides, preferably 60 to 120 nucleotides, upstream from the start site of transcription.

[0063] Public sequence: The term public sequence ," as used in the context of the instant application, refers to any sequence that has been deposited in a publicly accessible database. This term encompasses both amino acid and nucleotide sequences. Such sequences are publicly accessible, for example, on the BLAST databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast). The database at the NCBI GTP site utilizes gi" numbers assigned by NCBI as a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequence from various databases, including GenBank, EMBL, DBBJ, (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank).

[0064] Regulatory Sequence The term regulatory sequence," as used in the current invention, refers to any nucleotide sequence that influences transcription or translation initiation and rate, and stability and/or mobility of the transcript or polypeptide product. Regulatory sequences include, but are not limited to, promoters, promoter control elements, protein binding sequences, 5' and 3' UTRs, transcriptional start site, termination sequence, polyadenylation sequence, introns, certain sequences within a coding sequence, etc.

[0065] Related Sequences: Related sequences" refer to either a polypeptide or a nucleotide sequence that exhibits some degree of sequence similarity with a sequence described by the REF and SEQ tables.

[0066] Scaffold Attachment Region (SAR) As used herein, scaffold attachment region" is a DNA sequence that anchors chromatin to the nuclear matrix or scaffold to generate loop domains that can have either a transcriptionally active or inactive structure (Spiker and Thompson (1996) Plant Physiol. 110: 15-21).

[0067] Sequence-determined DNA fragments (SDFs) Sequence-determined DNA fragments" as used in the current invention are isolated sequences of genes, fragments of genes, intergenic regions or contiguous DNA from plant genomic DNA or cDNA or RNA the sequence of which has been determined.

[0068] Signal Peptide A signal peptide" as used in the current invention is an amino acid sequence that targets the protein for secretion, for transport to an intracellular compartment or organelle or for incorporation into a membrane. Signal peptides are indicated in the tables and a more detailed description located below.

[0069] Specific Promoter In the context of the current invention, specific promoters" refers to a subset of inducible promoters that have a high preference for being induced in a specific tissue or cell and/or at a specific time during development of an organism. By high preference" is meant at least 3-fold, preferably 5-fold, more preferably at least 10-fold still more preferably at least 20-fold, 50-fold or 100-fold increase in transcription in the desired tissue over the transcription in any other tissue. Typical examples of temporal and/or tissue specific promoters of plant origin that can be used with the polynucleotides of the present invention, are: PTA29, a promoter which is capable of driving gene transcription specifically in tapetum and only during anther development (Koltonow et al., *Plant Cell* 2:1201 (1990); RCc2 and RCc3, promoters that direct root-specific gene transcription in rice (Xu et al., *Plant Mol. Biol.* 27:237 (1995); TobRB27, a root-specific promoter from tobacco (Yamamoto et al., *Plant Cell* 3:371 (1991)). Examples of tissuespecific promoters under developmental control include promoters that initiate transcription only in certain tissues or organs, such as root, ovule, fruit, seeds, or flowers. Other suitable promoters include those from genes encoding storage proteins or the lipid body membrane protein, oleosin. A few root-specific promoters are noted above.

[0070] Stringency "Stringency" as used herein is a function of probe length, probe composition (G + C content), and salt concentration, organic solvent concentration, and temperature of hybridization or wash conditions. Stringency is typically compared by the parameter T_m, which is the temperature at which 50% of the complementary molecules in the hybridization are hybridized, in terms of a temperature differential from T_m. High stringency conditions are those providing a condition of T_m - 5°C to T_m - 10°C. Medium or moderate stringency conditions are those providing T_m - 20°C to T_m - 29°C. Low stringency conditions are those providing a condition of T_m - 40°C to T_m - 48°C. The relationship of hybridization conditions to T_m (in °C) is expressed in the mathematical equation

$$T_m = 81.5 - 16.6(\log_{10}[Na^+]) + 0.41(\%G+C) - (600/N) \quad (1)$$

where N is the length of the probe. This equation works well for probes 14 to 70 nucleotides in length that are identical to the target sequence. The equation below for T_m of DNA-DNA hybrids is useful for probes in the range of 50 to greater than 500 nucleotides, and for conditions that include an organic solvent (formamide).

$$T_m = 81.5+16.6 \log \{[Na^+]/(1+0.7[Na^+])\}+ 0.41(\%G+C)-500/L\ 0.63(\%formamide) \quad (2)$$

where L is the length of the probe in the hybrid. (P. Tijessen, Hybridization with Nucleic Acid Probes" in Laboratory Techniques in Biochemistry and Molecular Biology, P.C. vand der Vliet, ed., c. 1993 by Elsevier, Amsterdam.) The T_m of equation (2) is affected by the nature of the hybrid; for DNA-RNA hybrids T_m is 10-15°C higher than calculated, for RNA-RNA hybrids T_m is 20-25°C higher. Because the T_m decreases about 1 °C for each 1% decrease in homology when a long probe is used (Bonner et al., *J. Mol. Biol.* 81:123 (1973)), stringency conditions can be adjusted to favor detection of identical genes or related family members.

[0071] Equation (2) is derived assuming equilibrium and therefore, hybridizations according to the present invention are most preferably performed under conditions of probe excess and for sufficient time to achieve equilibrium. The time required to reach equilibrium can be shortened by inclusion of a hybridization accelerator such as dextran sulfate or another high volume polymer in the hybridization buffer.

[0072] Stringency can be controlled during the hybridization reaction or after hybridization has occurred by altering the salt and temperature conditions of the wash solutions used. The formulas shown above are equally valid when used to compute the stringency of a wash solution. Preferred wash solution stringencies lie within the ranges stated above; high stringency is 5-8°C below T_m, medium or moderate stringency is 26-29°C below T_m and low stringency is 45-48°C below T_m.

[0073] Substantially free of A composition containing A is substantially free of B when at least 85% by weight

of the total A+B in the composition is A. Preferably, A comprises at least about 90% by weight of the total of A+B in the composition, more preferably at least about 95% or even 99% by weight. For example, a plant gene or a DNA sequence can be considered substantially free of other plant genes or DNA sequences.

[0074] Translational start site In the context of the current invention, a translational start site" is usually an ATG in the cDNA transcript, more usually the first ATG. A single cDNA, however, may have multiple translational start sites.

[0075] Transcription start site Transcription start site" is used in the current invention to describe the point at which transcription is initiated. This point is typically located about 25 nucleotides downstream from a TFIID binding site, such as a TATA box. Transcription can initiate at one or more sites within the gene, and a single gene may have multiple transcriptional start sites, some of which may be specific for transcription in a particular cell-type or tissue.

[0076] Untranslated region (UTR) A UTR" is any contiguous series of nucleotide bases that is transcribed, but is not translated. These untranslated regions may be associated with particular functions such as increasing mRNA message stability. Examples of UTRs include, but are not limited to polyadenylation signals, terminations sequences, sequences located between the transcriptional start site and the first exon (5' UTR) and sequences located between the last exon and the end of the mRNA (3' UTR).

[0077] Variant: The term variant" is used herein to denote a polypeptide or protein or polynucleotide molecule that differs from others of its kind in some way. For example, polypeptide and protein variants can consist of changes in amino acid sequence and/or charge and/or post-translational modifications (such as glycosylation, etc).

DETAILED DESCRIPTION OF THE INVENTION

I. Polynucleotides

[0078] Exemplified SDFs of the invention represent fragments of the genome of corn, wheat, rice, soybean or *Arabidopsis* and/or represent mRNA expressed from that genome. The isolated nucleic acid of the invention also encompasses corresponding fragments of the genome and/or cDNA complement of other organisms as described in detail below.

[0079] Polynucleotides of the invention can be isolated from polynucleotide libraries using primers comprising sequence similar to those described by the REF and SEQ Tables. See, for example, the methods described in Sambrook et al., supra.

[0080] Alternatively, the polynucleotides of the invention can be produced by chemical synthesis. Such synthesis methods are described below.

[0081] It is contemplated that the nucleotide sequences presented herein may contain some small percentage of errors. These errors may arise in the normal course of determination of nucleotide sequences. Sequence errors can be corrected by obtaining seeds deposited under the accession numbers cited above, propagating them, isolating genomic DNA or appropriate mRNA from the resulting plants or seeds thereof, amplifying the relevant fragment of the genomic DNA or mRNA using primers having a sequence that flanks the erroneous sequence, and sequencing the amplification product.

I.A. Probes, Primers and Substrates

[0082] SDFs of the invention can be applied to substrates for use in array applications such as, but not limited to, assays of global gene expression, for example under varying conditions of development, growth conditions. The arrays can also be used in diagnostic or forensic methods (WO95/35505, US 5,445,943 and US 5,410,270).

[0083] Probes and primers of the instant invention will hybridize to a polynucleotide comprising a sequence in REF and SEQ TABLES 1 AND 2. Though many different nucleotide sequences can encode an amino acid sequence, the sequences of REF and SEQ TABLES 1 AND 2 are generally preferred for encoding polypeptides of the invention. However, the sequence of the probes and/or primers of the instant invention need not be identical to those in REF and SEQ TABLES 1 AND 2 or the complements thereof. For example, some variation in probe or primer sequence and/or length can allow additional family members to be detected, as well as orthologous genes and more taxonomically distant related sequences. Similarly, probes and/or primers of the invention can include additional nucleotides that serve as a label for detecting the formed duplex or for subsequent cloning purposes.

[0084] Probe length will vary depending on the application. For use as primers, probes are 12-40 nucleotides, preferably 18-30 nucleotides long. For use in mapping, probes are preferably 50 to 500 nucleotides, preferably 100-250 nucleotides long. For Southern hybridizations, probes as long as several kilobases can be used as explained below.

[0085] The probes and/or primers can be produced by synthetic procedures such as the triester method of Matteucci et al. *J. Am. Chem. Soc.* 103:3185(1981); or according to Urdea et al. *Proc. Natl. Acad.* 80:7461 (1981) or using commercially available automated oligonucleotide synthesizers.

I.B. Methods of Detection and Isolation

[0086] The polynucleotides of the invention can be utilized in a number of methods known to those skilled in the art as probes and/or primers to isolate and detect polynucleotides, including, without limitation: Southerns, Northerns, Branched DNA hybridization assays, polymerase chain reaction, and microarray assays, and variations thereof. Specific methods given by way of examples, and discussed below include:

Hybridization
Methods of Mapping
Southern Blotting
Isolating cDNA from Related Organisms
Isolating and/or Identifying Orthologous Genes.

Also, the nucleic acid molecules of the invention can used in other methods, such as high density oligonucleotide hybridizing assays, described, for example, in U.S. Pat. Nos. 6,004,753; 5,945,306; 5,945,287; 5,945,308; 5,919,686; 5,919,661; 5,919,627; 5,874,248; 5,871,973; 5,871,971; and 5,871,930; and PCT Pub. Nos. WO 9946380; WO 9933981; WO 9933870; WO 9931252; WO 9915658; WO 9906572; WO 9858052; WO 9958672; and WO 9810858.

B.1. Hybridization

[0087] The isolated SDFs of REF and SEQ TABLES 1 AND 2 of the present invention can be used as probes and/or primers for detection and/or isolation of related polynucleotide sequences through hybridization. Hybridization of one nucleic acid to another constitutes a physical property that defines the subject SDF of the invention and the identified related sequences. Also, such hybridization imposes structural limitations on the pair. A good general discussion of the factors for determining hybridization conditions is provided by Sambrook et al. ("Molecular Cloning, a Laboratory Manual, 2nd ed., c. 1989 by Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY; *see esp.*, chapters 11 and 12). Additional considerations and details of the physical chemistry of hybridization are provided by G.H. Keller and M.M. Manak DNA Probes", 2nd Ed. pp. 1-25, c. 1993 by Stockton Press, New York, NY.

[0088] Depending on the stringency of the conditions under which these probes and/or primers are used, polynucleotides exhibiting a wide range of similarity to those in REF and SEQ TABLES 1 AND 2 can be detected or isolated. When the practitioner wishes to examine the result of membrane hybridizations under a variety of stringencies, an efficient way to do so is to perform the hybridization under a low stringency condition, then to wash the hybridization membrane under increasingly stringent conditions.

[0089] When using SDFs to identify orthologous genes in other species, the practitioner will preferably adjust the amount of target DNA of each species so that, as nearly as is practical, the same number of genome equivalents are present for each species examined. This prevents faint signals from species having large genomes, and thus small numbers of genome equivalents per mass of DNA, from erroneously being interpreted as absence of the corresponding gene in the genome.

[0090] The probes and/or primers of the instant invention can also be used to detect or isolate nucleotides that are identical" to the probes or primers. Two nucleic acid sequences or polypeptides are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below.

[0091] Isolated polynucleotides within the scope of the invention also include allelic variants of the specific sequences presented in REF and SEQ TABLES 1 AND 2. The probes and/or primers of the invention can also be used to detect and/or isolate polynucleotides exhibiting at least 80% sequence identity with the sequences of REF and SEQ TABLES 1 AND 2 or fragments thereof.

[0092] With respect to nucleotide sequences, degeneracy of the genetic code provides the possibility to substitute at least one base of the base sequence of a gene with a different base without causing the amino acid sequence of the polypeptide produced from the gene to be changed. Hence, the DNA of the present invention may also have any base sequence that has been changed from a sequence in REF and SEQ TABLES 1 AND 2 by substitution in accordance with degeneracy of genetic code. References describing codon usage include: Carels *et al.*, *J. Mol. Evol.* 46: 45 (1998) and Fennoy *et al.*, *Nucl. Acids Res.* 21(23): 5294 (1993).

B.2. Mapping

[0093] The isolated SDF DNA of the invention can be used to create various types of genetic and physical maps of the genome of corn, Arabidopsis, soybean, rice, wheat, or other plants. Some SDFs may be absolutely associated with particular phenotypic traits, allowing construction of gross genetic maps. While not all SDFs will immediately be

[2347] The suspension culture cells are transformed with exogenous DNA as described by Z. Chen et al. *Plant Mol. Bio.* 36:163 (1998). Briefly, 4-days post-subculture cells are incubated with cell wall digestion solution containing 0.4 M sorbitol, 2% driselase, 5mM MES (2-[N-Morpholino] ethanesulfonic acid) pH 5.0 for 5 hours. The digested cells are pelleted gently at 60 xg for 5 min. and washed twice in W5 solution containing 154 mM NaCl, 5 mM KCl, 125 mM $CaCl_2$ and 5mM glucose, pH 6.0. The protoplasts are suspended in MC solution containing 5 mM MES, 20 mM $CaCl_2$, 0.5 M mannitol, pH 5.7 and the protoplast density is adjusted to about 4×10^6 protoplasts per ml.

[2348] 15-60 µg of plasmid DNA is mixed with 0.9 ml of protoplasts. The resulting suspension is mixed with 40% polyethylene glycol (MW 8000, PEG 8000), by gentle inversion a few times at room temperature for 5 to 25 min. Protoplast culture medium known in the art is added into the PEG-DNA-protoplast mixture. Protoplasts are incubated in the culture medium for 24 hour to 5 days and cell extracts can be used for assay of transient expression of the introduced gene. Alternatively, transformed cells can be used to produce transgenic callus, which in turn can be used to produce transgenic plants, by methods known in the art. See, for example, Nomura and Komamine, *Plt. Phys.* 79: 988-991 (1985), *Identification and Isolation of Single Cells that Produce Somatic Embryos in Carrot Suspension Cultures.*

[2349] The invention being thus described, it will be apparent to one of ordinary skill in the art that various modifications of the materials and methods for practicing the invention can be made. Such modifications are to be considered within the scope of the invention as defined by the following claims.

[2350] Each of the references from the patent and periodical literature cited herein is hereby expressly incorporated in its entirety by such citation.

Claims

1. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which encodes an amino acid sequence exhibiting at least 40% sequence identity to an amino acid sequence encoded by

   (a) a nucleotide sequence described in REF and/or SEQ Table 1 or 2 or a fragment thereof; or
   (b) a complement of a nucleotide sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof.

2. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which exhibits at least 65% sequence identity to

   (a) a nucleotide sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof; or
   (b) a complement of a nucleotide sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof.

3. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which exhibits at least 65% sequence identity to a gene comprising

   (a) a nucleotide sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof; or
   (b) a complement of a nucleotide sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof.

4. An isolated nucleic acid molecule which is the reverse of the isolated nucleotide sequence according to any one of claims 1-3, such that the reverse nucleotide sequence has a sequence order which is the reverse of the sequence order of said isolated nucleotide sequence according to any one of claims 1-3.

5. An isolated nucleic acid molecule comprising a nucleic acid capable of hybridizing to a nucleic acid having a sequence selected from the group consisting of:

   (a) a nucleotide sequence which is shown in REF and/or SEQ Table 1 or 2; and
   (b) a nucleotide sequence which is complementary to a nucleotide sequence shown in REF and/or SEQ Table 1 or 2;

   under conditions that permit formation of a nucleic acid duplex at a temperature from about 40°C and 48°C below the melting temperature of the nucleic acid duplex.

6. The nucleic acid molecule according to any one of claims 1-5, wherein said nucleic acid comprises an open reading frame.

7. The isolated nucleic acid molecule of any one of claims 1-5, wherein said nucleic acid is capable of functioning as a promoter, a 3' end termination sequence, an untranslated region (UTR), or as a regulatory sequence.

8. The isolated nucleic acid molecule of claim 7, wherein said nucleic acid is a promoter and comprises a sequence selected from the group consisting of a TATA box sequence, a CAAT box sequence, a motif of GCAATCG or any transcriptoin-factor binding sequence, and any combination thereof.

9. The isolated nucleic acid molecule of claim 7, wherein the nucleic acid sequence is a regulatory sequence which is capable of promoting seed-specific expression, embryo-specific expression, ovule-specific expression, tapetum-specific expression or root-specific expression of a sequence or any combination thereof.

10. A vector construct comprising a nucleic acid molecule according to any one of claims 1-9, wherein said nucleic acid molecule is heterologous to any element in said vector construct.

11. A vector construct according to claim 10 comprising:

(a) a first nucleic acid having a regulatory sequence capable of causing transcription and/or translation; and
(b) a second nucleic acid having the sequence of said isolated nucleic acid molecule according to any one of claims 1-4;

wherein said first and second nucleic acids are operably linked and wherein said second nucleic acid is heterologous to any element in said vector construct.

12. The vector construct according to claim 11, wherein said first nucleic acid is native to said second nucleic acid.

13. The vector construct according to claim 11, wherein said first nucleic acid is heterologous to said second nucleic acid.

14. A vector construct according to claim 10 comprising:

(c) a first nucleic acid having having the sequence of said isolated nucleic acid molecule according to claim 7; and
(d) a second nucleic acid;

wherein said first and second nucleic acids are operably linked and wherein said first nucleic acid is heterologous to any element in said vector construct.

15. The vector construct according to claim 14, wherein said first nucleic acid is native to said second nucleic acid.

16. The vector construct according to claim 14, wherein said first nucleic acid is heterologous to said second nucleic acid.

17. A host cell comprising an isolated nucleic acid molecule according to any one of claims 1-4, wherein said nucleic acid molecule is flanked by exogenous sequence.

18. A host cell comprising a vector construct of any one of claims 10-16.

19. An isolated polypeptide comprising an amino acid sequence

(a) exhibiting at least 40% sequence identity of an amino acid sequence encoded by a sequence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof; and
(b) capable of exhibiting at least one of the biological activities of the polypeptide encoded by said nucleotide seqence shown in REF and/or SEQ Table 1 or 2 or a fragment thereof.

20. The isolated polypeptide of claim 19, wherein said amino acid sequence exhibits at least 75% sequence identity to an amino acid sequence encoded by a sequence shown in SEQ Table 1 or 2 or a fragment thereof.

21. The isolated polypeptide of claim 19, wherein said amino acid sequence exhibits at least 85% sequence identity

to an amino acid sequence encoded by a sequence shown in SEQ Table 1 or 2 or a fragment thereof.

22. The isolated polypeptide of claim 19, wherein said amino acid sequence exhibits at least 90% sequence identity to an amino acid sequence encoded by a sequence shown in SEQ Table 1 or 2 or a fragment thereof.

23. An antibody capable of binding the isolated polypeptide of any one of claims 19-22.

24. A method of introducing an isolated nucleic acid into a host cell comprising:

(a) providing an isolated nucleic acid molecule according to any one of claims 1-4; and
(b) contacting said isolated nucleic with said host cell under conditions that permit insertion of said nucleic acid into said host cell.

25. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to any one of claims 10-16.

26. A method of modulating transcription and/or translation of a nucleic acid in a host cell comprising:

(a) providing the host cell of claim 24 or 25; and
(b) culturing said host cell under conditions that permit transcription or translation.

27. A method for detecting a nucleic acid in a sample which comprises:

(a) providing an isolated nucleic acid molecule according to any one of claims 1-5;
(b) contacting said isolated nucleic acid molecule with a sample under conditions which permit a comparison of the sequence of said isolated nucleic acid molecule with the sequence of DNA in said sample; and
(c) analyzing the result of said comparison.

28. The method according to claim 27, wherein said isolated nucleic acid molecule and said sample are contacted under conditions which permit the formation of a duplex between complementary nucleic acid sequences.

29. A plant or cell of a plant which comprises a nucleic acid molecule according to any one of claims 1-4 which is exogenous to said plant or plant cell.

30. A plant or cell of a plant which comprises a nucleic acid molecule according to any one of claims 1-4, wherein said nucleic acid molecule is heterologous to said plant or said cell of a plant.

31. A plant or cell of a plant which has been transformed with a nucleic acid molecule according to any one of claims 1-4.

32. A plant of cell of a plant which comprises a vector construct according to any one of claims 10-16.

33. A plant of cell of a plant which has been transformed with a vector construct according to any one of claims 10-16.

34. A plant which has been regenerated from a plant cell according to any one of claims 29-33.